**DOCUMENT RESUME**

ED 319 774 TM 014 987

AUTHOR Harwell, Michael R.; And Others
TITLE Summarizing Monte Carlo Results in Methodological Research: The Oneway Fixed-Effects ANOVA Case.
PUB DATE Apr 90
NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Analysis of Variance; *Data Interpretation; Educational Research; *Meta Analysis; *Methods Research; *Monte Carlo Methods; Robustness (Statistics); Sample Size; Testing Problems
IDENTIFIERS *F Test; *Type I Errors

ABSTRACT

Concern over the validity of statistical tests performed on data that may not satisfy underlying assumptions has prompted methodological researchers to perform Monte Carlo studies for frequently used tests. Unfortunately, these studies appear to have had little impact on methodological practice. One reason is the lack of an overarching framework to guide the interpretation of Monte Carlo studies for the same test. Another is the impressionistic nature of these studies, which can lead different readers to different conclusions. These shortcomings can be addressed using quantitative methods of research synthesis (e.g., meta-analysis) to summarize the results of Monte Carlo studies for a statistical test. In this paper, these methods are applied to a sample of Monte Carlo studies of the F-test in the oneway fixed-effects analysis of variance (ANOVA) model. The present analyses were based on Monte Carlo studies reported in 21 out of 30 journal articles. The results provide empirical support for the robustness of the Type I error rate of the F-test to certain assumption violations. However, the Type I error rate of the F-test was noticeably affected by unequal variances, even when sample sizes were equal. Recommendations for using this test when certain assumptions are violated are made. Four data tables and one bar graph are included. (Author/TJH)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED319774

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Michael R. Harwell
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

# Summarizing Monte Carlo Results in Methodological Research: The Oneway Fixed-Effects ANOVA Case

Michael R. Harwell

William S. Hayes

Corley C. Olds

Elaine N. Rubinstein

University of Pittsburgh

Paper presented at the Annual Meeting of the American Educational Research Association, Boston, 1990.



TM014987

2

BEST COPY AVAILABLE

## Abstract

Concern over the validity of statistical tests performed on data that may not satisfy underlying assumptions has prompted methodological researchers to perform Monte Carlo studies for frequently used tests. Unfortunately, these studies appear to have had little impact on methodological practice. One reason is the lack of an overarching framework to guide the interpretation of Monte Carlo studies for the same test. Another is the impressionistic nature of these studies, which can lead different readers to different conclusions. These shortcomings can be addressed using quantitative methods of research synthesis (e.g., meta-analysis) to summarize the results of Monte Carlo studies for a statistical test. In this paper, these methods are applied to a sample of Monte Carlo studies of the F-test in the oneway fixed-effects ANOVA model. The results provide empirical support for the robustness of the type I error rate of the F-test to certain assumption violations. However, the type I error rate of the F-test was noticeably affected by unequal variances, even when sample sizes were equal. Recommendations for using this test when certain assumptions are violated are made.

# Summarizing Monte Carlo Results in Methodological Research: The Oneway Fixed-Effects ANOVA Case

## Introduction

An ongoing concern of quantitative methodologists is the validity of statistical tests performed on data that may not satisfy underlying assumptions (e.g., normality of a population score distribution). These concerns have been heightened by recent work suggesting that the bulk of educational and psychological data are at least moderately and sometimes strikingly nonnormal (Micceri, 1989). Micceri's work is evidence of the usefulness of statistical tests which are insensitive to assumption violations, i.e., whose type I error properties are not deleteriously affected. Tests which are insensitive to assumption violations are considered to be robust; tests which are not robust are less useful.

A large number of MC studies of particular statistical tests are available in the methodological research literature. Unfortunately, these results lack an overarching framework to guide their interpretation. In addition, the impressionistic nature of MC results makes it possible for different readers to reach different conclusions. These shortcomings can be addressed by using quantitative methods of research synthesis (e.g., meta-analysis) (Harwell, 1990): the goal is to quantitatively summarize the results of MC studies for a statistical test in a way that generates guidelines for using that test under specific assumption violations. This would also permit the results of previous statistical analyses using that test to be evaluated.

The purpose of this paper is to apply the meta-analytic framework

illustrated in Harwell (1990) to summarize the results of a sample of MC studies of the F-test in the oneway fixed-effects ANOVA model. The paper is organized following the framework illustrated in Cooper (1982) and used in Harwell (1990). Only the type I error case is examined in this paper.

First, previous attempts to summarize MC studies of the F-test in the oneway fixed-effects ANOVA model are briefly reviewed. The need to complement qualitative summaries of MC studies with quantitative methods is emphasized. Next, data collection procedures and issues are discussed. Then, data evaluation procedures which are used to ensure accurate coding and data entry are discussed. Finally, the MC data is analyzed and the results interpreted. These results inform methodological practice by generating guidelines for using the F-test in the oneway fixed-effects ANOVA model under specific assumption violations.

## Problem Formulation

A number of MC studies of the F-test are available in the methodological research literature. However, previous attempts to summarize these results have been narrative in nature and have lacked an overarching framework to guide their interpretation. These shortcomings can be addressed using methods conceptualized by Glass (1976), who suggested using standardized mean differences (i.e., effect magnitudes) as a way of summarizing study results. In the present context, the empirical proportions of rejections (i.e., empirical type I errors) serve

as effect magnitudes (EMs) (Harwell, 1990). The goal of these methods is to produce an empirical network of MC results which will generate guidelines for the use of the F-test under specific assumption violations.

The F-test was selected for two reasons. First, comparing the meta-analytic results to known theoretical and empirical results for this popular test permits the usefulness of the meta-analytic methods to be evaluated. Second, these methods will be used to investigate the effect of heterogeneous variances on the F-test when sample sizes are equal. Recent MC evidence (e.g., Tomarken & Serlin, 1987) cast doubt on the oft-cited conclusion of Glass, Peckham, and Sanders (1972) that, in the presence of equal samples, there is a "very slight effect on $\alpha$ [the nominal type I error rate], which is seldom disturbed by more than a few hundredths".

## Data Collection

### Selection of Studies

A population of MC studies of the F-test in the oneway fixed-effects ANOVA model was identified by searching the ERIC data base, Dissertation Abstracts International, and the Current Index to Statistics. Key words used to locate relevant studies follow: ANOVA, distribution-free, Kruskal-Wallis, Monte Carlo, nonnormality, nonparametric, power, ranks, robustness, simulation, t-tests, Type I error rate, Wilcoxon, and Welch. The literature search yielded approximately thirty journal articles and four dissertations that appeared to be accessible.

Searching large data bases does not ensure that all relevant studies

will be identified. For example, MC results reported in unpublished technical reports and master's theses are likely to be underepresented or missed completely. Under these conditions, the MC studies included in the meta-analysis may differ in some important way from those not included. The nature of MC studies, however, makes it probable that the potentially nonrandom sample of MC studies of the F-test is representative of the specified population.

The small number of accessible studies yielded by the literature search led to the decision to use every available study in the meta-analysis. Note that one or more study selection biases may be introduced if the identified population of studies are not representative of the entire population of MC studies of the F-test in the oneway fixed-effects ANOVA model (Harwell, 1990).

The present analyses were based upon MC studies reported in twenty-one of the thirty journal articles. These articles are listed in appendix A. The data reported in the remaining articles and the dissertations are not yet available for statistical analysis. Hence the conclusions in this paper are preliminary and could change with the inclusion of the remaining MC studies.

Next, the twenty-one MC studies were screened for serious methodological flaws. The fact that all twenty-one studies were published in refereed journals provides some protection. In addition, each study was examined for inconsistent or unusual procedures and results using the following criteria: a) how the data were generated (e.g., random number generator used), b) evidence of the success of the data generation (e.g., skewness and kurtosis statistics computed for the

simulated data), and c) the pattern of empirical type I error results when underlying assumptions of the F-test were satisfied (e.g., whether the empirical type I error rate of the F-test converged toward the nominal value as sample size increased if all assumptions are satisfied). No irregularities were noted and thus all twenty-one studies were judged to be methodologically sound.

Coding of Outcome and Explanatory Variables

The outcome variable for the meta-analysis was type I error rate. This variable was coded directly. Only results associated with a nominal level of .05 were coded. Several characteristics of the MC studies were coded as explanatory (i.e., predictor) variables. They are listed below:

(1) type of population score distribution

normal ($\gamma_1 = 0$, $\gamma_2 = 0$)
uniform ($\gamma_1 = 0$, $\gamma_2 = -1.12$)
double-exponential ($\gamma_1 = 0$, $\gamma_2 = 3$)
log-normal ($\gamma_1$, $\gamma_2$ depend on the parameters used)
Cauchy ($\gamma_1 = 0$, $\gamma_2$ undefined)
exponential ($\gamma_1 = 2$, $\gamma_2 = 6$)
logistic ($\gamma_1 = 0$, $\gamma_2 = 4.2$)
t ($\gamma_1 = 0$, $\gamma_2 = \nu/(\nu-4)$, $\nu$ = error degrees of freedom)
mixed normal ($\gamma_1$, $\gamma_2$ defined for each application)
other [the other category includes the binomial distribution ($\gamma_1$, $\gamma 2$ depend on the proportion of successes and sample size) and Poisson distribution ($\gamma_1$, $\gamma_2$ depend on parameter specified)

(2) number of groups

(3) total sample size

(4) ratio of largest to smallest sample size

1 = 1 (sample sizes equal)
2 = > 1 and <1.25
3 = > 1.25 and < 1.5
4 = > 1.5 and < 1.75
5 = > 1.75 and < 2.0
6 = > 2.0 and < 3
7 = > 3 and < 5
8 = > 5

(5) ratio of largest to smallest variance

1 = 1 (all variances equal)
2 = > 1 and < 2
3 = > 2 and < 3
4 = > 3 and < 5
5 = > 5 and < 8
6 = > 8

(6) pairing of sample size and variance

1 = positively correlated (e.g., large variances paired with large sample sizes)
2 = negatively correlated (e.g., large variances paired with smaller samples)
3 = other

(7) number of samples (replications)

The population score distribution information was captured by coding skewness ($\gamma_1$) and kurtosis ($\gamma_2$) values (Kendall & Stuart, 1977, Vol. I, pp. 187-189). The $\gamma_1$ and $\gamma_2$ indices for the unimodal but skewed and kurtic lognormal distribution depend on the selected parameters and hence two MC studies employing a lognormal distribution may be examining quite different distributions. Similarly, the $\gamma_1$ and $\gamma_2$ indices for the binomial and Poisson distributions depend on the parameters specified in the MC study and if this information was not reported, which was often the case, these indices could not be coded. The kurtosis associated with

a Cauchy distribution could not be coded since the variance theoretically does not exist.

The selection of ranges for coding the pattern of sample sizes and the pattern of variances was guided by conditions reported in the sample of MC studies. The number of replications variable is the number of randomly generated samples upon which the empirical type I error values are based. This variable was coded since it is related to the magnitude of sampling error of the empirical proportions of rejections.

## Data Evaluation

### Accuracy of Coding and Data Entry

A three phase process was used to ensure that the characteristics of each MC study were accurately coded and correctly entered into a computer data file in preparation for statistical analysis. In an initial training phase, two of the twenty-one MC studies were reviewed and coded by all four authors. The structure of one of these studies was relatively simple and the other was more complex. Coding forms based on the above coding scheme were completed for each study by each author. The completed coding forms were then compared. Instances of uncertainty or disagreement over particular characteristics of a MC study (e.g., how sample sizes and variances were paired) were resolved by group consensus. Information from this training phase was used to modify the coding forms.

In the next phase, eight of the twenty-one MC studies were equally divided among two teams of coders, each made up of two of the authors. The members of a team independently reviewed and coded each article

assigned to them using the modified coding forms. Members of a team then compared their results and attempted to resolve discrepancies among themselves. Only a few instances of inconsistent coding were encountered. Each of the remaining MC studies was coded by one of the authors.

In the third phase, the coded MC data were entered into a computer data file and then checked for accuracy. The twenty-one MC studies generated approximately 553 lines of data (i.e., 553 EMs). The size and complexity of the data set virtually guaranteed errors in data entry. Two strategies were used to detect and correct data entry errors. First, a computer printout of the entire data file was scanned in order to detect obvious errors, e.g., type I error values falling outside an expected range. Second, a comprehensive check was carried out by randomly assigning the twenty-one articles to the four authors, having each author read the articles assigned to them, and check the coded data. Errors in coding and data entry detected in this fashion were then corrected.

## Data Analysis and Interpretation

The goal of quantitatively summarizing MC results for a particular statistical test is to construct a statistical model that explains the

Table 1 +@

Summary Statistics for Quantitative Variables
for the Sample of Monte Carlo Studies

| Variable | Cases | Mean | Median | Stdev | Minimum | Maximum |
|---|---|---|---|---|---|---|
| TYPEI | 553 | .059 | .050 | .039 | .004 | .309 |
| SKEW | 1056 | 1.08 | 0 | 1.78 | 0 | 6.19 |
| KURT | 1056 | 11.14 | 0 | 28.4 | -3.75 | 110.9 |
| TOTALN | 1225 | 45.4 | 32 | 43.5 | 8 | 750 |
| REPS1 | 1070 | 4289.6 | 2000 | 4110.5 | 400 | 10,000 |

+ Cases = number of MC cases, Stdev = standard deviation.

*@ Table 1 results include power values which were not considered in the inferential analyses later*

behavior of the statistical test as a function of study characteristics (Harwell, 1990). Recall that available analytic and empirical evidence of the behavior of the F-test will be directly compared against the meta-analytic results. This will provide evidence about the usefulness of the proposed methods. The relationship between heterogeneous variances and type I error when sample sizes are equal will also be investigated.

## Descriptive Analyses

The first stage of the data analysis was descriptive in nature. Statistics were computed for a variety of quantitative and qualitative variables. Summary information on the sample of twenty-one MC studies is given in Tables 1 and 2. The variables in these tables represent empirical type I error values (TYPEI), skewness (SKEW), kurtosis (KURT), total sample size (TOTALN), number of replications for the type I error case (REPS1), number of replications for the power case (REPS2), number of groups (NUMGRPS), ratio of largest to smallest sample sizes (SAMPLE), ratio of largest to smallest variances (VARIANCE), and pairing of sample

Table 2 + @

Summary Statistics for Qualtitative Variables
for the Sample of Monte Carlo Studies

Type of Population Score Distribution

| Catagory | Frequency | % |
|---|---|---|
| Normal | 450 | 36.7 |
| Uniform | 30 | 2.4 |
| Dbl. Exponential | 20 | 1.6 |
| Log-normal | 133 | 10.9 |
| Cauchy | 20 | 1.6 |
| Exponential | 182 | 14.9 |
| Logistic | 16 | 1.3 |
| t | 15 | 1.2 |
| Mixed-normal | 67 | 5.5 |
| Other | 292 | 23.8 |
| Total | 1225 | 100 |

Number of Groups

| Catagory | Frequency | % |
|---|---|---|
| 2 | 442 | 36.1 |
| 3 | 194 | 15.8 |
| 4 | 520 | 42.4 |
| 6 | 9 | .7 |
| 8 | 60 | 4.9 |
| Total | 1225 | 100 |

Ratio of Largest/Smallest Sample Size

| Catagory | Frequency | % |
|---|---|---|
| Equal | 763 | 62.3 |
| 1-1.25 | 0 | 0.0 |
| 1.25-1.5 | 49 | 4.0 |
| 1.5-1.75 | 0 | 0.0 |
| 1.75-2 | 11 | .9 |
| 2-3 | 243 | 19.8 |
| 3-5 | 138 | 12.9 |
| >5 | 1 | .1 |
| Total | 1225 | 100 |

Pairing of Sample Size/Variance

| Catagory | Frequency | % |
|---|---|---|
| Other | 1005 | 82 |
| Pos. Corr. | 124 | 10.1 |
| Neg. Corr. | 96 | 7.8 |
| Total | 1225 | 100 |

Ratio of Largest/Smallest Variance

| Catagory | Frequency | % |
|---|---|---|
| Equal | 878 | 71.7 |
| 1-2 | 119 | 9.7 |
| 2-3 | 42 | 3.4 |
| 3-5 | 87 | 7.1 |
| 5-8 | 39 | 3.2 |
| >8 | 60 | 4.9 |
| Total | 1225 | 100 |

+ Dbl. exponential = double exponential, Pos. Corr. = positively correlated, Neg. Corr. = negatively correlated.size and variance (PAIRING).

*@ Table 2 results incude power cases which were not considered in the inferential analyses*

The results in Table 1 indicate that the average type I error rate across the sample of N = 553 type I error values was quite close to the nominal value. A plot of the empirical TYPEI values appears in Figure 1. This distribution is noticeably skewed. Another interesting statistic in Table 1 is the difference between the minimum and maximum number of replications. This difference suggests results of varying precision. Table 2 contains summary statistics for qualitative variables.

Quantitative Analyses

To construct and evaluate explanatory models, a fixed-effects regression model was fitted to the empirical type I error values (see Hedges & Olkin, 1985, p.169). The fixed-effects regression models were of the form

$$p_k = X_1 \beta_1 + X_2 \beta_2 + \ldots + X_{KT} \beta_T \ , \ k=1,2,\ldots,K \qquad (1)$$

where $p_k$ is the k(th) EM which depends on a set of T fixed explanatory variables $X_{KT}$, and $\beta_T$ is a regression coefficient that captures the relationship between the t(th) predictor variable and the k(th) EM (see Harwell, 1990). In the present context, the empirical type I errors served as the $p_k$ and the coded characteristics of the MC studies as the $X_{KT}$.

Specified explanatory models were fitted to the EMs and a test of the relationship between the set of T predictor variables and the $p_k$ was performed using the weighted sum of squares due to regression statistic

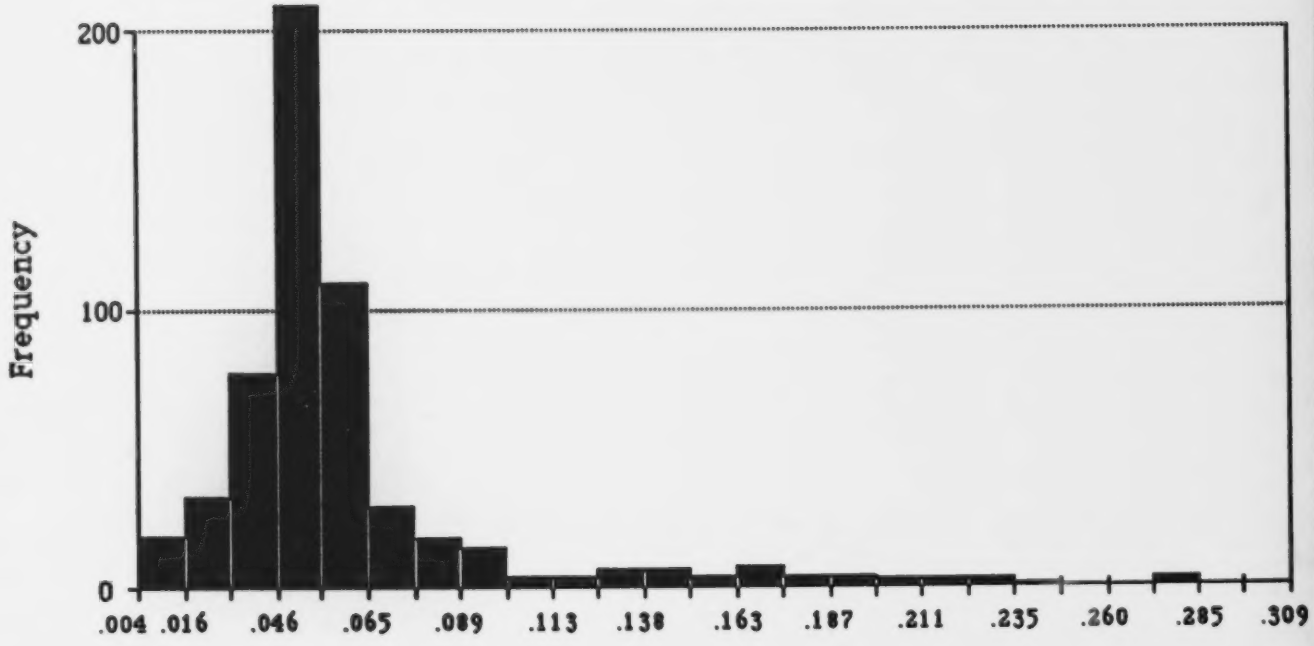
Frequency
200
100
0
.004
.016
.046
.065
.089
.113
.138
.163
.187
.211
.235
.260
.285
.309
TYPEI
(Empirical Type I error rate)

*FIGURE 1*

$Q_R$ given in Hedges and Olkin (1985, p. 171). Under the hypothesis of no relationship between the set of explanatory variables and the outcome variable, $Q_R$ is approximately distributed as a chi-square with T degrees of freedom. The squared multiple correlation coefficient was used as an index of the explanatory power of a model. A test of model misspecification (i.e., whether all of the explanatory variables needed to explain variation in the $p_k$ are in the model) was performed using the $Q_E$ statistic, also given in Hedges and Olkin (1985, p. 173). Under the hypothesis of no model misspecification, $Q_E$ is approximately distributed as a chi-square with K-T-1 degrees of freedom. All tests used an error rate of .05. Listwise deletion of missing data reduced the number of cases used in the analyses.

Six explanatory models were investigated for the type I error case using the SPSSX (1983) computer program. Each model is discussed below. A summary of the results of the regression analyses appears in Table 3. Examination of the residuals of each of the models indicated no noticeable departures from normality. Note that all of the $Q_R$ and $Q_E$ statistics in Table 3 are significant at $p < .001$ and are often quite large. Despite the misspecification of all of the models, the multiple $R^2$ statistic appeared to be a useful index of the explanatory power of a model.

Model 1

Model 1 investigated the relationship between type I errors and the predictor variables SKEW, KURT, VARIANCE, TOTALN, NUMGRPS, SAMPLE, and REPS1:

Table 3 +

| MODEL | T | Cases | $Q_R$* | $Q_E$* | $R^2$ | $R^2_{adj}$ |
|---|---|---|---|---|---|---|
| 1a | 7 | 416 | 2833.6 | 26181.6 | .10 | .08 |
| 1b | 59 | 416 | 6786.9 | 22228.3 | .23 | .11 |
| 2a | 5 | 416 | 2075.3 | 26939.9 | .07 | .06 |
| 2b | 7 | 416 | 2833.6 | 26181.6 | .10 | .08 |
| 3a | 6 | 416 | 2714.9 | 26300.3 | .09 | .08 |
| 3b | 7 | 416 | 2833.6 | 26181.6 | .10 | .08 |
| 4a | 7 | 416 | 2833.6 | 26181.6 | .10 | .08 |
| 4b | 8 | 416 | 2872.9 | 26142.3 | .10 | .08 |
| 5a | 7 | 149 | 2583.03 | 22429.5 | .10 | .06 |
| 5b | 8 | 149 | 17909.3 | 7103.2 | .72 | .70 |
| 5c | 7 | 76 | 4076.0 | 3179.2 | .56 | .52 |
| 5d | 7 | 73 | 777.6 | 237.6 | .77 | .74 |
| 6a | 5 | 195 | 1479.2 | 2440.2 | .38 | .36 |
| 6b | 6 | 195 | 2658.9 | 1260.5 | .68 | .67 |

+ T = number of predictors, Cases = number of cases, $Q_R$ is the weighted sum of squares due to regression statistic, $Q_E$ is a statistic testing model misspecification, * means significant at $p < .001$, $R^2$ is the squared correlation between the set of predictors and the outcome variabes, and $R^2_{adj}$ is $R^2$ adjusted for the number of predictors (see Marascuilo & Serlin, 1988, p. 661).

<u>model 1a</u>

$$TYPEI = SKEW\ \beta_1 + KURT\ \beta_2 + VARIANCE\ \beta_3 + TOTALN\ \beta_4 + NUMGRPS\ \beta_5 + SAMPLE\beta_6 + REPS1\ \beta_7$$

<u>model 1b</u>

$$TYPEI = SKEW\ \beta_1 + KURT\ \beta_2 + VARIANCE\ \beta_3 + TOTALN\ \beta_4 + NUMGRPS\ \beta_5 + SAMPLE\beta_6 + REPS1\ \beta_7 + 52 \text{ predictors representing two-variable-at-a-time and three-variable-at-a-time interaction effects}$$

In model 1a seven predictor variables were fitted to the TYPEI values. The results in Table 3 indicate that there is a statistically significant relationship between the set of seven predictor variables and TYPEI. However, the $R^2_{adj}$ value of .08 suggests that the model

possesses little explanatory power. This modest relationship supports the commonly held notion that the type I error rate of the F-test in the oneway fixed-effects ANOVA model is robust. In the present context, robustness would be indicated by a weak relationship between aset of predictor variables and the outcome variable TYPEI. Model 1b was used to investigate the relationship between interactions of predictor variables and TYPEI with the effects of the seven original predictor variables held constant. Fifty-two predictors representing almost all possible two-variable-at-a-time and three-variable-at-a-time interactions among the T = 7 predictors in 1a were entered after the seven original predictors. Collinearity problems prohibited four of the interactions from being entered into the model. Although the increase in the $Q_R$ statistic between models 1a and 1b ($Q_{R\ 1b} - Q_{R\ 1a} = 3953.3$) is statistically significant, the relatively small difference in the adjusted $R^2$s (.12) suggests that the addition of the interaction effects only slightly increased the explanatory power of the model. On the whole, the results of model 1b suggest that the type I error rate of the F-test is relatively insensitive to multiple assumption violations.

Model 2

Model 2 was used to investigate the effect of the shape of the population score distribution, as captured with skewness and kurtosis indices, on type I errors. The models investigated were:

<u>model 2a</u>

TYPEI = VARIANCE $\beta_1$ + TOTALN $\beta_2$ + NUMGRPS $\beta_3$ + SAMPLE $\beta_4$ + REPS1 $\beta_5$

model 2b

$$TYPEI = VARIANCE\ \beta_1 + TOTALN\ \beta_2 + NUMGRPS\ \beta_3 + SAMPLE\ \beta_4 + REPS1\ \beta_5 + SKEW\ \beta_6 + KURT\ \beta_7$$

Comparing the results for models 2a and 2b indicates that the $Q_{R\ 2b} - Q_{R\ 2a}$ difference was significant; however, the difference in the $R^2$s suggests that type of population score distribution had little to do with explaining variation in the type I errors. This result supports the perception that the type I error rate of the F-test is robust to departures from the assumption of normality of a population score distribution.

Model 3

The effect of the number of replications variable on type I errors was investigated in model 3. The models were:

model 3a

$$TYPEI = VARIANCE\ \beta_1 + TOTALN\ \beta_2 + NUMGRPS\ \beta_3 + SAMPLE\ \beta_4 + SKEW\ \beta_5 + KURT\ \beta_6$$

model 3b

$$TYPEI = VARIANCE\ \beta_1 + TOTALN\ \beta_2 + NUMGRPS\ \beta_3 + SAMPLE\ \beta_4 + SKEW\ \beta_5 + KURT\beta_6 + REPS1\ \beta_7$$

The results in Table 3 indicate that, with the other predictors held constant, number of replications had a negligable impact on type I errors ($R^2_{adj\ 3b} - R^2_{adj\ 3a} = 0$).

Model 4

In model 4 the possibility of a quadratic relationship between type

I error and total sample size was investigated. The rationale was that, other factors held constant, as sample size increases the type I error rate should converge toward its nominal value, but there may be a point beyond which larger samples contribute little to this convergence. The models were:

model 4a

$$TYPEI = VARIANCE\ \beta_1 + TOTALN\ \beta_2 + NUMGRPS\ \beta_3 + SAMPLE\ \beta_4 + SKEW\ \beta_5 + KURT\beta_6 + REPS1\ \beta_7$$

model 4b

$$TYPEI = VARIANCE\ \beta_1 + TOTALN\ \beta_2 + NUMGRPS\ \beta_3 + SAMPLE\ \beta_4 + SKEW\ \beta_5 + KURT\beta_6 + REPS1\ \beta_7 + TOTALN^2\ \beta_8$$

The results in Table 3 suggests that there is no quadratic relationship between sample size and type I errors.

Model 5

The relationship between pairing unequal sample sizes and variances and type I errors was investigated in models 5a-5d. Theoretical and empirical work suggests that the meta-analysis should detect a strong relationship between the set of predictor variables (including PAIRING) and type I error. Models 5a and 5b were:

model 5a

$$TYPEI = SKEW\ \beta_1 + KURT\ \beta_2 + VARIANCE\ \beta_3 + TOTALN\ \beta_4 + NUMGRPS\ \beta_5 + SAMPLE\ \beta_6 + REPS1\ \beta_7$$

model 5b

$$TYPEI = SKEW\ \beta_1 + KURT\ \beta_2 + VARIANCE\ \beta_3 + TOTALN\ \beta_4 + NUMGRPS\ \beta_5 + SAMPLE\ \beta_6 + REPS1\ \beta_7 + PAIRING\ \beta_8$$

The results of these analyses, reported in Table 3, provide strong evidence of the relationship between type I error and pairing. Model 5a produces an $R^2_{adj} = .06$ which is similar to that of model 1a. Note, however, that model 5 analyses are restricted to MC results examining the effects of pairing and thus are based on a smaller sample. Model 5b includes the pairing variable and produces an $R^2_{adj} = .70$. The $R^2_{adj\ 5b} - R^2_{adj\ 5a} = .64$ and $Q_{R\ 5b} - Q_{R\ 5a} = 15326.3$ differences suggests a strong relationship between type I errors and the pairing of unequal sample sizes and variances, with the effects of the other predictor variables held constant. These results are consistent with theoretical and previous empirical evidence (Glass et al., 1972).

Specific evidence about the role of sample size and variance pairings were examined through models 5c and 5d. Model 5c investigated the relationship between type I error and the predictors skewness, kurtosis, number of groups, total sample size, and number of replications but was restricted to MC data in which sample sizes and variances were positively correlated, e.g., smaller samples paired with smaller variances; model 5d investigated this relationship when samples and variances were negatively correlated, e.g., larger samples paired with larger variances. The models were:

<u>model 5c</u> (sample sizes/variances positively correlated)

$$TYPEI = SKEW\ \beta_1 + KURT\ \beta_2 + TOTALN\ \beta_3 + NUMGRPS\ \beta_4 + REPS1\ \beta_5$$

<u>model 5d</u> (sample sizes/variances negatively correlated)

$$TYPEI = SKEW\ \beta_1 + KURT\ \beta_2 + TOTALN\ \beta_3 + NUMGRPS\ \beta_4 + REPS1\ \beta_5$$

The results for models 5c and 5d in Table 3 suggest a strong relationship between type I error and the explanatory models for positive and negative pairing of sample sizes and variances.

Model 6

Model 6 investigated the relationship between type I error and heterogeneous variances when sample sizes are equal. All of the data used in these analyses are based on equal sample sizes. The models were:

<u>model 6a</u> (equal sample sizes)

$$\text{TYPEI} = \text{SKEW}\ \beta_1 + \text{KURT}\ \beta_2 + + \text{TOTALN}\ \beta_3 + \text{NUMGRPS}\ \beta_4 + \text{REPS1}\ \beta_5$$

<u>model 6b</u> (equal sample sizes)

$$\text{TYPEI} = \text{SKEW}\ \beta_1 + \text{KURT}\ \beta_2 + \text{TOTALN}\ \beta_3 + \text{NUMGRPS}\ \beta_4 + \text{REPS1}\ \beta_5 + \text{VARIANCE}\ \beta_6$$

The adjusted $R^2$ for model 6b (.67) and the difference $R^2_{\text{adj 6b}} - R^2_{\text{adj 6a}} = .35$ suggests a strong relationship between variance inequality and type I error even though sample sizes are equal. Further evidence of this effect is provided by examining the type I error means for the variance condition variable when sample sizes are equal. This information is presented in Table 4. Variance ratios greater than 2 produce a noticeably inflated type I error rate, a pattern that is exacerbated as

Table 4

Average Type I Error Rates By Variance Ratios
For Equal Sample Sizes

| | VARIANCE | | | | | |
|---|---|---|---|---|---|---|
| | Equal | 1-2 | 2-3 | 3-5 | 5-8 | > 8 |
| Mean | .046 | .051 | .060 | .064 | .064 | .079 |
| N of cases | 144 | 7 | 15 | 15 | 6 | 23 |

the variance ratio increases. This pattern persists even if the variance condition is restricted to ratios < 5. In this case, analysis of a model identical to 5b (not given) produced an $R^2_{adj} = .54$. These results sugests that equal samples provide little protection against inflated type I error rates when variances are heterogeneous.

## Conclusions

The results of the meta-analysis suggest the following conclusions:

1. The type I error rate of the F-test is insensitive to type of population score distribution and relatively insensitive to combinations of violations of assumptions.
2. There was no relationship between the number of replications and type I errors, despite the large differences in these values across studies.
3. There is no quadratic relationship between sample size and type I errors.

4. There is a strong relationship between inverse pairing of sample sizes and variances and type I error. This extends to the case in which sample sizes and variances are positively paired.
5. There is a moderate relationship between the set of predictor variables and type I error when sample sizes are equal; for unequal sample sizes there is only a weak relationship.
6. Equal sample sizes provide little protection against inflated type I error rates when variances are heterogeneous. This pattern is present for variance ratios < 5.

## Summary

The application of quantitative methods of research synthesis to summarize Monte Carlo results shows great promise for informing methodological practice. Construction of an empirical framework of Monte Carlo studies of a statistical test should result in guidelines for the appropriate use of particular statistical tests under specific assumption violations. This will also permit previous statistical analyses to be evaluated considering these guidelines.

The present results suggest that meta-analytic methods can usefully be applied to summarizing Monte Carlo results of particular statistical tests. The results support the commonly help perception of the robustness of the type I error rate of the oneway fixed-effects ANOVA F-test for a variety of conditions. Nonnormal population score distributions, different sample sizes, numbers of groups, and unequal sample sizes had little effect on type I errors. However, the results of the meta-analysis provide new evidence that researchers should not rely

on equal sample sizes to neutralize the effects of heterogeneous variances. Under these conditions, the likely result is an inflated type I error rate.

The next step in the process of deriving guidelines for using the F-test when assumptions are violated is to tease out more specific information from the explanatory models identified as being correlated with type I errors. The goal would be to identify conditions (e.g., variance inequality and sample size) associated with specific type I error values. This requires a more sophisticated methodology (e.g., response surface methodology). The same process should also be used to investigate the relationship between various explanatory models and the power of the F-test.

## References


Blair, C. (1981). A reaction to "Consequences of failure to meet assumptions underlying fixed effects analysis of variance and covariance." Review of Educational Research, 51, 499-507.

Box, G.E.P. (1953). Non-normality and tests on variances. Biometrika, 40, 318-335.

Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3-8.

Glass, G.V., McGaw, B., & Smith, M.L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.

Games, P.A. (1978). Testing for homogeneity of variance and covariance. Review of Educational Research, 42, 237-288.

Harwell, M.R. (1990). Summarizing Monte Carlo results in methodological research. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, April.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. New York: Harcourt, Brace, Jovanovich:

Hsu, P. L. (1938). Contribution to the theory of "student's t-test as applied to the problem of two samples. Statistical Research Memoirs, 2, 1-24.

Ito, P.K. (1980). Robustness of ANOVA and MANOVA procedures. In P.R. Krishnaiah (ed.), Handbook of Statistics, Vol. I, 199-236. New York: North Holland Publishing Company.

Kendall, M., & Stuart, A. (1977). The advanced theory of statistics. New York: Macmillan.

Marascuilo, L.A., & Serlin, R.C. (1988). **Statistical methods for the social and behavioral sciences.** New York: Freeman.

Micceri, T. (1989). The unicorn, the normal distribution, and other improbable creatures. **Psychological Bulletin, 105,** 156-166.

Rosenthal, R. (1978). Combining results of independent studies. **Psychological Bulletin, 85,** 185-193.

SPSSX, Inc. (1983). New York: McGraw-Hill.

Tomarkin, A., & Serlin, R.C. (1986). A comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. **Psychological Bulletin, 99,** 90-99.

## Articles Used in the Meta-Analysis


Blair, R. C., Higgins J.J., & Smitley W.D.S. (1980). On the relative power of the U and t tests. British Journal of Mathematical and Statistical Psychology, 33, 114-120.

Boehnke, K. (1984). F- and H-test assumptions revisited. Educational and Psychological Measurement, 44, 609-617.

Boneau, C.A. (1960). The effects of violations of assumptions underlying the t test. Psychological Bulletin, 57, 49-64.

Budescu, D.V., & Appelbaum, M.I. (1981). Variance stabilizing transformations and the power of the F-test. Journal of Educational Statistics, 6, 55-74.

Clinch, J.J., & Keselman, H.J. (1982). Parametric alternatives to the analysis of variance. Journal of Educational Statistics, 1, 207-214.

Dijkstra, J.B., & Werter, P.S.P.J. (1981). Testing the equality of several means when the population variances are unequal. Communications in Statistics: Simulation and Computation, 10, 557-569.

Donaldson, T.S. (1968). Robustness of the F-test to errors of both kinds and the correlation between the numerator and denominator of the F-ratio. Journal of the American Statistical Association, , 660-676.

Neave, H.R., & Granger, W.J. (1968). A Monte Carlo study comparing various two-sample tests for differences in means. **Technometrics**, 10, 509-522.

Olejnik, S. (1987). Conditional ANOVA for mean differences when population variances are unknown. **Journal of Experimental Education**, , 141-148.

Penfield, D.A., & Koffler, S.L. (1985). A power study of selected nonparametric K-sample tests. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, March.

Rasmussen, J.L. (1985). An evaluation of parametric and non-parametric tests on modified and non-modified data. **British Journal of Mathematical and Statistical Psychology**, 39, 213-220.

Rogan, J.C., & Keselman, H.J. (1977). Is the ANOVA F-test robust of variance heterogeneity when sample sizes are equal?: An investigation via a coefficient of variation. **American Educational Research Journal**, 14, 493-498.

Tomarken, A.J., & Serlin, R.C. (1986). A comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. **Psychological Bulletin**,

Wilcox, R.R., Charlin, V.L., & Thompson, K.L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W and $F^*$ statistics. **Communication in Statistics: Simulation and Computation**, 15, 933-943.

Feir-Walsh, B.J., & Toothaker, L.E. (1974). An empirical comparison of the ANOVA F-test, normal scores test and Kruskal-Wallis test under violation of assumptions. Educational and Psychological Measurement, 34, 789-799.

Games, P.A., & Lucas, P.A. (1966). The analysis of variance of independent groups on non-normal and normally transformed data. Educational and Psychological Measurement, 26, 311-327.

Kohr, R.L., & Games, P.A. (1974). Robustness of the analysis of variance, the Welch procedure and a Box procedure to heterogeneous variances. Journal of Experimental Education, 43, 61-69.

Levine, D.W., & Dunlap, W.P. (1982). Power of the F test with skewed data: should one transform or not? Psychological Bulletin, 92, 272-280.

Lin, L.I., & Sanford, R.L. (1983). The robustness of the likelihood ratio test, the nonparametric rank sum test, and F-ratio tests when the populations are from the negative binomial family. Communication in Statistics: Simulation and Computation, 12, 523-539.

McSweeney, M., & Penfield, D. (1969). The normal scores test for the c-sample problem. British Journal of Mathematical and Statistical Psychology, 22, 177-192.

Nath, R., & Duran, B.S. (1981). The rank transform in the two-sample location problem. Communication in Statistics: Simulation and Computation, 10, 383-394.